Determining t in t-closeness using Multiple Sensitive Attributes

Debaditya Roy

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela – 769 008, India

Determining t in t-closeness using Multiple Sensitive Attributes

Dissertation submitted in

May 2013

to the department of

Computer Science and Engineering

of

National Institute of Technology Rourkela

in partial fulfillment of the requirements

for the degree of

Master of Technology

by

Debaditya Roy

(Roll no. 211CS1050)

under the supervision of

Dr. Sanjay Kumar Jena

Department of Computer Science and Engineering

National Institute of Technology Rourkela

Rourkela – 769 008, India

To my Parents

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, India. www.nitrkl.ac.in

Dr. Sanjay Kumar Jena
Professor

May 29, 2013

Certificate

This is to certify that the work in the thesis entitled *Determining t in t-closeness using Multiple Sensitive Attributes* by *Debaditya Roy*, bearing roll number 211CS1050, is a record of original research work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of *Master of Technology* in *Computer Science and Engineering* with specialization in *Computer Science*. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Sanjay Kumar Jena

Acknowledgement

This dissertation, though an individual work, has benefited in various ways from several people. Whilst it would be simple to name them all, it would not be easy to thank them enough.

The enthusiastic guidance and support of *Prof. Sanjay Kumar Jena* inspired me to stretch beyond my limits. His profound insight has guided my thinking to improve the final product. My solemnest gratefulness to him.

My humble acknowledgment to *Prof. K. S. Babu* for his constructive criticism during entire span of research.

It further gives me a deep sense of pride to have received the tutelage of *Prof. S. K. Rath*,*Prof. B. Majhi* and *Prof. A. K. Turuk*.

Overwhelming thanks to all members of the Department of Computer Science and Engineering, NIT Rourkela for their encouragement and co-operation throughout.

Many thanks to my fellow research colleagues and classmates at *Advanced Database Engineering Lab*. It gives me immense happiness to be graduating with such an energetic batch of students.

Finally, my heartfelt thanks to my family for their unconditional love and support. Words fail me to express my gratitude to my beloved parents, who egged me on every step of the way.

Debaditya Roy

Abstract

Many government agencies and other organizations often need to publish microdata, e.g., medical data or census data, for research and other purposes. Typically, such data is stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories. (1) Attributes that clearly identify individuals. These are known as explicit identifiers and include Social Security Number, Address, and Name, and so on. (2) Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers (QI), and may include, e.g., Zip-code, Birthdate, and Gender. (3) Attributes that are considered sensitive, such as Disease and Salary are known as Sensitive Attributes (SA). When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed.

Therefore, the objective is to limit the disclosure risk to an acceptable level while maximizing the utility.This can be achieved by anonymizing the data before release.Models like *k-anonymity*(to prevent *linkage attacks*), *l-diversity*(to prevent *skewness attacks*), *t-closeness*(to prevent *background knowledge attacks*) etc. have been proposed over the years which are collectively known as Privacy Preserving Data Publishing models.

Here, a novel way in determining t and applying *t-closeness* for multiple sensitive attributes is presented. The only information required beforehand is the partitioning classes of Sensitive Attribute(s). Since, t-closeness is an *NP-Hard* problem, so knowing thee value of t greatly reduces the time required for anonymizing with various values of t . The rationale of using the measure of determining t is discussed with conclusive proof and speedup achieved is also shown.

***Keywords*:Privacy Preserving Data Mining, Privacy Preserving Data Publishing, t-closeness, Multiple Sensitive Attributes**

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Many government agencies and other organizations often need to publish microdata i.e. data pertaining to an individual's medical health, telvision viewership or any other behavioral data which can be mined to get overall statistics about the population as a whole.The example of such microdata medical data or census data which are published for research and other purposes. Typically, such data is stored in a table, and each record (row) corresponds to one individual.

Each record has a number of attributes, which can be divided into the following three categories. (1) Attributes that clearly identify individuals. These are known as *explicit identifiers* and include Social Security Number, Address, and Name, and so on. (2) Attributes whose values when taken together can potentially identify an individual. These are known as *quasi-identifiers (QI)* [1], and may include, e.g., Zip-code, Birthdate, and Gender. QIs are context-dependent and may vary according to the data dissemination forums available.(3) Attributes that are considered sensitive, such as Disease and Salary are known as *Sensitive Attributes (SA)* [1]. These are the attributes the user wants hidden from public view in general or not to be directly identified with him in specific.

When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed as shown in Figure 1.1. While the released table gives useful information to researchers, it presents *disclosure risk* to the individuals

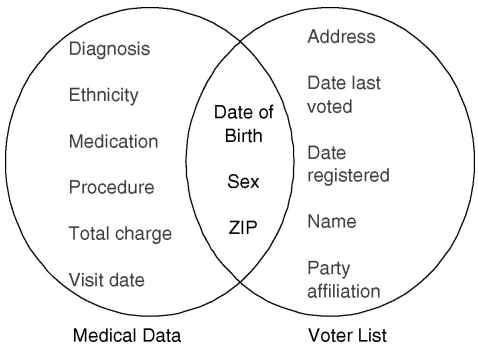

Figure 1.1: Linking to reidentify owner [Sweeney 2002]

whose data are in the table. Therefore, the objective is to limit the disclosure risk to an acceptable level while maximizing the utility. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly-available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. This is also known as *background knowledge attack*.

A common anonymization approach is generalization, which replaces quasi-identifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of QI values.This leads to the definition of an *equivalence class*. An *equivalence class* of an anonymized table and of the sensitive attribute is defined as a set of records that have the same values for all the QIs. Another is suppression, which suppresses the value of an attribute if that value causes the overall k-anonymity or any other privacy measure to fail. But the suppression is minimized using the maximum suppression count or percentage.

Models like k-anonymity [2, 3], l-diversity [4, 5], t-closeness [6] etc. have been proposed over the years which are collectively known as Privacy Preserving Data Publishing models. In [1] k-anonymity was introduced as the property that each

record is indistinguishable with at least k-1 other records with respect to the quasi-identifier. In [4] a new notion of privacy was introduced, called l-diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least "well represented" values.

In [6], t-closeness was proposed that formalizes the idea of global background knowledge by requiring that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). Furthermore, the t-closeness approach tends to be more effective than many other privacy-preserving data mining methods for the case of numeric attributes. However, all the above methods were applied on datasets where only a single sensitive attribute was considered.

In the following thesis a novel way of calculating t in t-closeness using multiple sensitive attributes is presented. The partitioning of sensitive attributes (SA) into classes and finding their *Niche Overlap* [7] gives a way to find t. As t-closeness is an NP-Hard problem [8], execution time can be saved if application of any t-closeness algorithm is carried out for one value of t. Once this is accomplished, the verification is done using the results obtained by checking with the various metrics of Information Loss.

1.1 Motivation

The main issue with t-closeness is its specificity of application. It is totally dependent upon the dataset in question because of its dependence on k-anonymity for complete execution. This dependence causes high overlead in deployment of any anonymization algorithm involving t-closeness.

The main emphasis of this thesis was on reducing the effective time in deploying anonymization while also reducing the dependence on specific datasets for every step of the process starting from the defintion of Domain Generalization Hierarchy

to Sensitive Attribute selection.

1.2 Problem Statement

This thesis aims at providing with an approach to find t for any given dataset before anonymization thereby making the job of anonymization faster and efficient. The problem with implementing t-closeness without having any fixed t is that any t-closeness problem is an NP-Hard Problem [8] and solving any instance of it for a practical dataset(generally with millions of records) takes high execution time. Moreover, the cross-verification to check with different values of t to determine the least possible value is also time consuming. The proposed scheme is based on finding Niche Overlap between Multiple Sensitive Attributes and then using that to determine t.

1.3 Thesis Layout

The thesis is organised into five chapters including this one. They systematically deal with the problem at hand and the reason for taking the approach and also detailed explanation of the proposed scheme.

Chapter 2: Literature Review This chapter deals with the existing literature on Privacy Preserving Data Publishing with a focus on t-closness and Mutliple Sensitive Attributes. Also Niche Overlap is discussed at length for clear understanding of its role in the proposed scheme. For the sake of completeness, the various Privacy Attacks and Information Loss Metrics are also discussed.

Chapter 3: Proposed Scheme This chapter details the procedure used to find the formula for calculating t. Also the choice of suitable Sensitive Attributes is discussed.Finally, a detailed inference follows regarding the correctness of the formula derived.

Chapter 4: Experiment and Results This chapter exhibit the results obtained after determining Niche Overlap [7] on the ADULT dataset. Further, the behaviour of various Information Loss metrics

Chapter 5: Conclusion This chapter sheds light on the contribution of the thesis meanwhile discussing the various results obtained in the previous chapter and their implications. The various extensions of this work is mentioned in Future Scope.

Chapter 2

Literature Review

2.1 k-anonymity

The notion of k-anonymity was given by Samarati and Sweeney [1–3, 9]. They stated that if one record in the table has some value *qid* then atleast *k-1* other record also have the same value *qid*. So, the minimum group size on *QID* is k. A table follwing this requirement is *k-anonymous*.

To understand the concept of k-anonymity the following tables are presented. The Table 2.1 containing three sensitive attributes *ZIP Code, Age* and *Disease* refers to the original patients table given in [6]. The 3-anonymous table is presented in Table 2.2. Similarly, other k-anonymous tables can be constructed.

Table 2.1: Original Patients' Table

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

The notion of k-anonymity suffers from *Homogenity attack* and *Background Knowledge attack*. In the first case, if there is no diversity in the equivalence class and if the adversary if able to point out the equivalence class itself then the sensitive attribute being same for all will disclose the sensitive information of the victim. This is *homogenity attack*. Next, if the adversary has some backgorund information on the victim then if there is a probabilistic case where two or more sensitive attributes are possible, the backgorund knowledge may help in deciding. This is *background knowledge attack*.

Table 2.2: A 3-anonymous version of Table 2.1

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

2.2 l-diversity

The definition of *l-diversity* according to [4] is as follows.

> *A q^*-block is l-diverse if contains at least l well-represented values for the sensitive attribute S. A table is l-diverse if every q^*-block is l-diverse.*

To understand l-diversity the follwing tables are presented. The Original Salary/Disease table is presented in Table 2.3 containing the attributes *ZIP Code, Age, Salary* and *Disease*originally referenced from [10]. The 3-diverse version (on *Disease*) is given in Table 2.4.

However, l-diversity is vulnerable to both *Skewness attack* and *Similiarity attack*. Suppose we have equal number of positive and negative records in an equivalence class of size 100. This follows 2-diversity but it is vulnerable to the fact that anyone in the class has 50% possibility of being positive rather than 1%. This is *skewness attack*. Next if the SA values in an equivalence class are distinct but semantically same i.e. all represent a range of values having similar properties for e.g. low income then we face *similarity attack*.

Table 2.3: Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric
2	47602	22	4K	gastritis
3	47678	27	5K	stomach
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach

Table 2.4: A 3-diverse version of Table 2.3

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach

2.3 t-closeness

Privacy gain is measured by the information gain of an observer [6]. Before seeing the released table the observer has some prior belief B0 about the sensitive attribute of an individual. If the observer sees a table that is completely generalized (quasi identifier are either removed or generalized equivalently) then the belief becomes B_1 which is influenced by Q, the distribution of sensitive attribute of the table. Now, when he sees the actually released table, by knowing the quasi identifier(s) of the table the observer is able to learn about P, the distribution of the sensitive attribute of the table, his belief changes to B_2.

In order for the public information to be Q, do not limiting the gain between B_0 and B_1 is not prevented. Rather, the distance between P and Q is limited and the closer they are B_2 does not vary much from B_1 and the gained knowledge from the released table is quite less. This results in maintaining privacy for the participants of the data. The table presented here

Table 2.5: Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

2.4 Distance Measures

Till now, two distance metrics for calculating t in t-closeness have been used.

2.4.1 Earth Movers Distance

The Earth Movers Distance (EMD) [11] measures the distance between two distributions, in this case the distance between the distribution of SA in an equivalence class and the overall distribution of that SA in the table or dataset. The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. This distance is different for both numerical and categorical attributes. The forms describing them are given as follows. For numerical attributes, let $r_i = p_i - q_i$,$(i = 1, 2, ..., m)$, then the EMD between P and Q can be calculated as:

$$\begin{aligned} D[P,Q] &= \frac{1}{m-1}(|r_1| + |r_1 + r_2| + ... + |r1 + r_2 + ... + rm|) \\ &= \frac{1}{m-1}\sum_{i=1}^{i=m}\left|\sum_{j=1}^{j=i} r_j\right| \end{aligned} \tag{2.1}$$

For categorical attributes the *Hierarchical Distance* is given as follows :
The distance between two values of a categorical attribute is based on the minimum level to which these two values are generalized to the same value according to the domain hierarchy.

Given a domain hierarchy and two distributions P and Q, we define the extra of a leaf node that corresponds to element i, to be $p_i - q_i$, and the extra of an internal node N to be the sum of extras of leaf nodes below N. This extra function can be defined recursively as:

$$extra(N) = \textstyle\sum_{C \epsilon Child(N)} extra(C) \text{ [N is non-leaf node]} \tag{2.2}$$

where $Child(N)$ is the set of all leaf nodes below node N.The extra function has the property that the sum of extra values for nodes at the same level is 0. Further define two other functions for *internal nodes*:

$$pos\_extra(N) = \sum_{C \epsilon Child(N) \wedge extra(C) > 0} |extra(C)| \tag{2.3}$$

We use $cost(N)$ to denote the cost of movings between $N's$ children branches. An optimal flow moves exactly $extra(N)$ in/out of the subtree rooted at N. Suppose that pos $extra(N) > neg\_extra$, then $extra(N) = pos\_extra(N) - neg\_extra(N)$ and $extra(N)$ needs to move out. (This cost is counted in the cost of $N's$ parent node.) In addition, one has to move neg_extra among the children nodes to even out all children branches; thus,

$$cost(N) = \frac{height(N)}{H} min(pos\_extra(N), neg\_extra(N)) \tag{2.4}$$

Then the earth mover's distance can be written as:

$$D[P, Q] = \sum_N cost(N) \tag{2.5}$$

where N is a non-leaf node.

2.4.2 Hellinger distance

Hellinger distance [12] as the distance measure for calculating t was proposed in [13] is calculated for two distributions P and Q as follows:

$$H(P,Q) = \sqrt{(1 - BC(P,Q))} \tag{2.6}$$

where $BC(P,Q)$ is Bhattacharya's Coefficient [14].

2.5 Multiple Sensitive Attributes (MSA)

The problem of MSA was first tackled in [15] based on k-anonymity [2] and l-diversity [4], where it was determined that generalization is not the solution in this case. Further, a framework known as Decomposition was given in [16] which was based on l-diversity [3] was given to tackle the MSA in any given table. In [17] , an improved framework known as Decomposition+ was given with implementations on real-life scenarios.

Another model for MSA was given in [18], which was based on (n, t) closeness [10] stating the limitations of l-diversity.

2.6 Information Loss Metrics

To measure the loss in data quality occurring when we generalize the microdata the following the metrics were proposed.

2.6.1 Discernibility Metric

The Discernibility Metric (DM) [19], measures the cardinality of the equivalence class. It assigns a penalty to each tuple based on how many tuples in the transformed dataset are indistinguishable from it. Let t be a tuple from the original table T, and let $G_{T^*}(t)$ be the set of tuples in an anonymized table T^* indistinguishable from t or the set of tuples in T^* equivalent to the anonymized

value of t. Then, DM is defined as follows:

$$DM(T^*) = \sum_{t \epsilon T} |G_{T^*}(t)| \tag{2.7}$$

2.6.2 Precision

The precision [19] of a generalization scheme is measured as:
1 - the average height of a generalization (measured over all cells).
The precision is *1* if there is no generalization and is *0* if all values are generalized.

2.6.3 Non-Uniform Entropy

Given a set of tuples S and the class labels cls involved in S, the entropy [20] is defined as:

$$H(S) = -\sum_{c \epsilon cls} \frac{freq(S,c)}{|S|} \times log_2 \frac{freq(S,c)}{|S|} \tag{2.8}$$

where $freq(S, c)$ is the number of tuples containing class c in S.

These metrics together along with others give an idea about the available information utility after generalization.

2.7 Niche Overlap

The concept of Niche Overlap is given in [7]. It can be calculated for all kinds of data but here the concentration is on *Categorical Data*.

Suppose there are K categories (e.g. of habitat), all assumed to be equally available to species i. The proportional use of category k by species i is written p_{ik}, assuming $\sum_{k=1}^{K} p_{ik} = 1$. In K categories, the niche overlap between species i and j

is defined as:

$$NO_{ij} = \sum_{k=1}^{K} min(p_{ik}p_{jk}) \tag{2.9}$$

The Niche Overlap(NO)is shown as in Figure 2.1 between two attributes *Education* and *Relationship*. The summation of the minimum area between the each of the pair of columns is considered to be the *NO*.

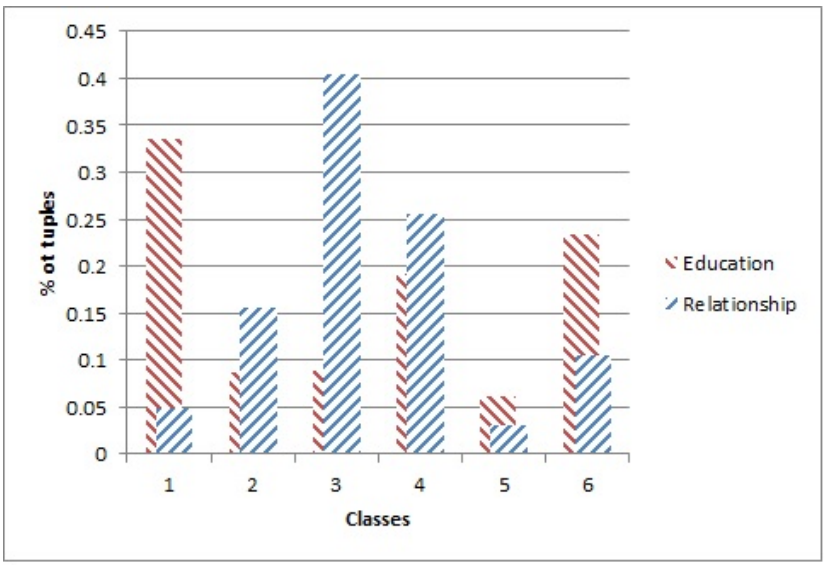

Figure 2.1: Niche Overlap between Education and Relationship

Chapter 3

Proposed Scheme

3.1 Determining Niche Overlap and *t*

While determining the optimal value of t the following problems are encountered. (1) In known literature, there is no mention of any method for determining t. All the authors [21] suggest is to match the heuristics based on earlier experiments to verify the results and this involves a lot of randomness in the experiments. (2) If the optimum value of t has to be determined using the utility vs. privacy curve it is not possible do so because of the inherent nature of the curve i.e. diverging.

So, instead a method based on the partitioning of sensitive attributes into classes is employed. The sensitive attributes to be considered for this exercise can vary according to the necessity of the data disseminating body in question [1].

After partitioning as shown for education class in Figure 3.1., all the categorical sensitive attributes are coded to numerical values and all the continuous numerical sensitive attributes which were previously categorized as intervals or classes are also coded. Once this is achieved the area of overlap between the various sensitive attributes was determined. The least overlap between all these attributes is considered as the Niche Overlap between all the sensitive attributes, *NO*. The algorithm 1 details the process.

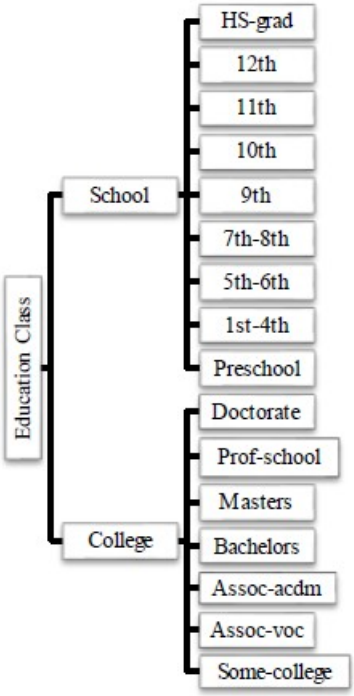

Figure 3.1: Partitioning of Education Class

$$t = 1 - NO \tag{3.1}$$

Once the value of t according to the equation 3.1, the value obtained can be used as an upper bound for applying anonymization on the dataset T for which it was calculated. This reduces the effort to anonymize the dataset for the values of t greater than the obtained t since they will not give minimum data quality degredation. The reason for this and the justification for our scheme is given in the next section.

Algorithm 1 Obtain Niche Overlap *(NO)* between *SAs*

1. Clean the dataset , removing any missing or unknown values.
2. For the chosen SAs S_i, construct partritioning classes $p_{i1}$to p_{in}(if mismatch occurs in number of classes then generalize using DGH).
3. For each SA S_i which is categorical encode the classes with numeric values.
4. For each pair of SAs , $S_i$and $S_j$where $i \neq j$, do
   (a) For each partitioning class $p_{ik}$and p_{jk}, k=1 to n,do
      i. Find the minimum overlap i.e. $m_k = min(|p_{ik}|, |p_{jk}|)$ where $|p|$ represents the percentage of tuples under partitioning class p.
      ii. NO_{new}=$NO_{new} + m_k$
   (b) Update NO if $NO_{new} \leq NO$(initially set to 1).
   (c) Assign the m_k to either $S_i$or S_j and remove the other from the list.
5. Obtain the final value of NO.

3.2 Detailed Inference

In this section, the detailed inference to the formula that is stated as equation 3.1 in the earlier section to get the t-value is provided. The definition of t-closeness [6] states:

> *An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.*

From the above definition two things can be inferred.First, the lower the value of t i.e. $t \rightarrow 0$; the more diverse the original data in dataset T is and the equivalence

class is required to be as close to original data as possible to give the required anonymization. Secondly, the higher the value of t i.e. $t \rightarrow 1$, the less diverse the original data in dataset T is and the equivalence class is required to be as different as possible from the original data to give the appropriate anonymization. This follows from the definition of EMD as well which determines t as the distance between the two distributions: *equivalence class* and *overall table*.

From these two inferences it can be said that if it is possible to capture the diversity of two or more sensitive attribute by the diversity of their equivalence classes and then it can be used to get a common diversity measure by comparing the Niche Overlap [7] between the two distributions.

Since, t-closeness captures the similarity between the distributions by definition we need to subtract the Niche Overlap value from the total area under the curve which in this case is equal to *1*. Hence, equation 3.1 used to calculate t is indeed correct.

3.3 Regarding choice of Multiple Sensitive Attributes

This is an area which is generally not defined well in literature, as it varies from dataset to dataset and also implementation to implementation. The main idea from [16] and [17] is that choosing MSA is a big task in executing their algorithms. However, this can be simplified if the inspection into the *DGH* of the every SA up for selection is made beforehand.

The way this works is if the number of partitioning classes is not equal for the SAs in question then it is better to choose some other combinations. In case, this is not possible and anonymization has to be done beforehand the anonymization should be done for the SA with most uniform distribution so as not to affect the outcome of the proposed scheme. The most uniformly distributed SA will retain its properties on anonymization thereby unhindering further calculations of *NO*.

Chapter 4

Experiment and Results

The experiment was carried out on the ADULT dataset [22]. All the papers on t-closeness [6, 10] etc. have taken this dataset itself for publishing results. The ADULT dataset was first cleaned removing missing attributes totalling the number of records to $30,162$. Next, nine (9) out of the fourteen(14) attributes were chosen namely,$Age, Final-Weight, MartialStatus, Race, Gender$, $Work-class, Education, HoursperWeek$ and $Relationship$. The first four were deemed as QIDs and $Workclass$ was chosen as the Primary Sensitive Attribute and the remaining were deemed as MSA. The distribution of SAs is shown in Figures 4.1,4.2,4.3 and 4.4.

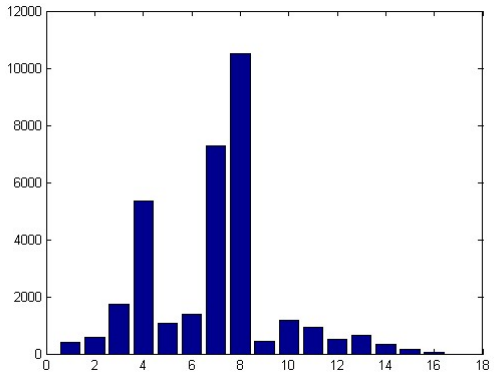

Figure 4.1: education distribution

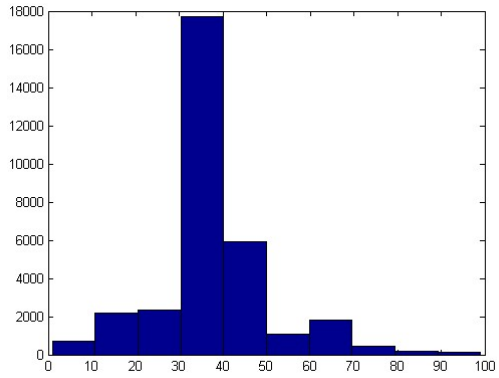

Figure 4.2: hours-per-week distribution

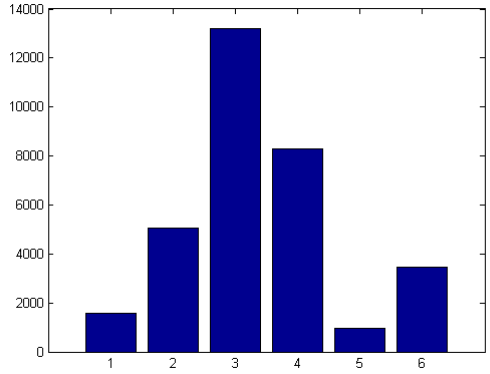

Figure 4.3: relationship distribution

4.1 Determining Niche Overlap

The area of overlap calculated in terms of Niche Overlap [7] which gives the area of overlap between two discrete distributions between the attributes provided for any SA S_i $\sum_{k=1}^{K} p_{ik} = 1$ where K gives all possible discrete values S_i takes. One such Niche Overlap is calculated between Relationship and Education shown in Table 4.2.

The detailed Niche Overlap values are shown in Table 4.1 and the scattered plots

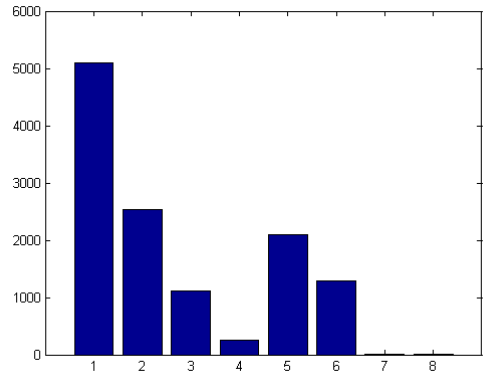

Figure 4.4: workclass distribution

Table 4.1: Niche Overlap Areas between Sensitive Attributes

Sensitive Attribute Group	Niche Overlap
Hours-per-week-relationship-workclass	89% or 0.89
Hours-per-week-relationship-education	85% or 0.85
Hours-per-week-relationship-workclass	79% or 0.79
Education-workclass-relationship	30% or 0.30

are given in Figures 4.5,4.6,4.7 and 4.8.The figures are plotted and the analysis is done using Matlab.

4.2 Determining t

The trio of education class-relationship status-workclass gives such low value of overlap because of the high mismatch in the number of classes of the participating attributes. Hence, it can be safely considered that 79% or 0.79 is the minimum overlap value i.e. $NO = 0.79$ and according to equation 3.1 the value of $t = 0.21$.

Taking the upper bound on t obtained above the verification was done as follows.

Table 4.2: Niche Overlap Areas between Sensitive Attributes

SA /class encoding	1	2	3	4	5	6	Total
relationship	0.0481	0.1556	0.4051	0.2550	0.0301	0.1058	1
education	0.3351	0.0862	0.0890	0.1930	0.0622	0.2341	1
Overlap	0.0481	0.0862	0.0890	0.1930	0.0301	0.1058	0.5524

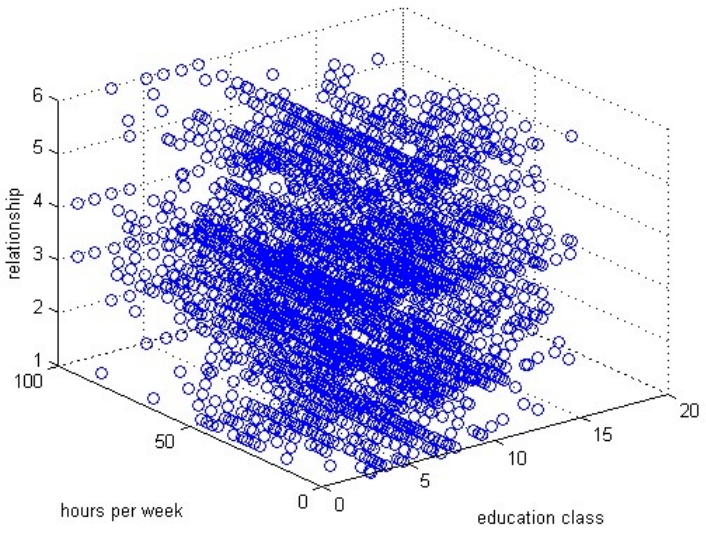

Figure 4.5: Scatter plot of hours-per-week vs relationship vs education class(ADULT)

4.3 Verification of t-values

The ARX-Flash Anonymization Toolbox [23] was used to check various Information Loss metrics like $Precision$,$DiscernibilityMetric$ and $Non-Uniform\ Entropy$ along with varying values of t in the range *(0.10,0.15,0.20,0.25,0.30)* and k in the range *(2-35)* with hierarchial EMD. The results obtained are shown in Figures 4.9,4.10,4.11.

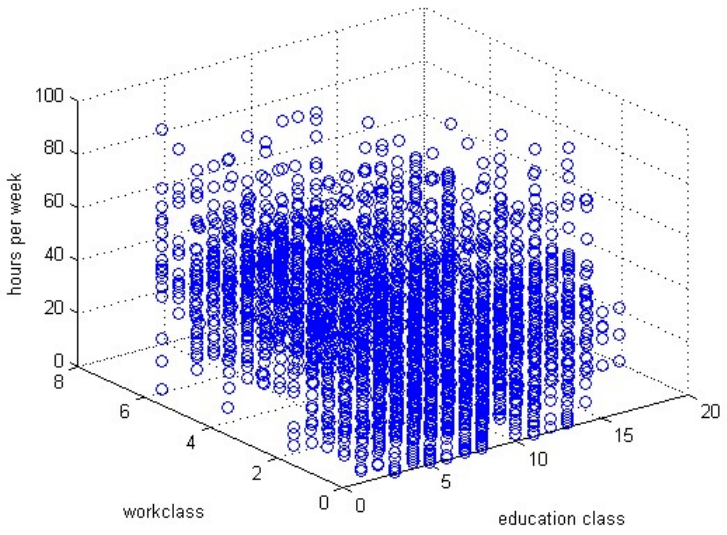

Figure 4.6: Scatter plot of hours-per-week vs workclass vs education class(ADULT)

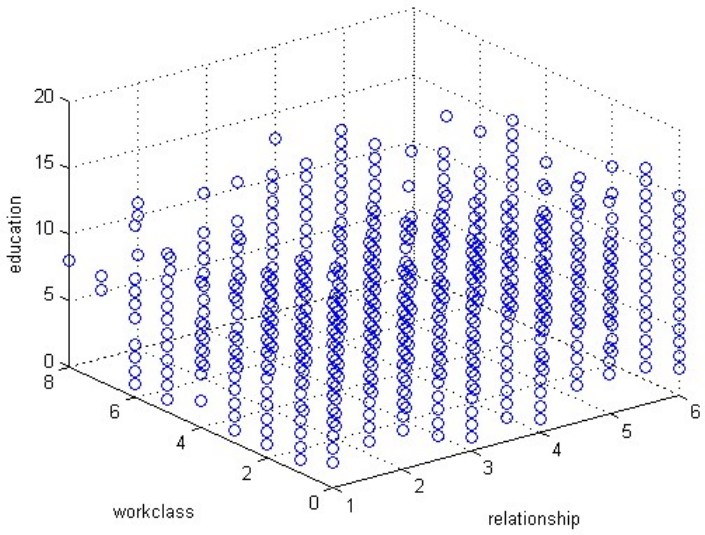

Figure 4.7: Scatter plot of education class vs. workclass vs. relationship(ADULT)

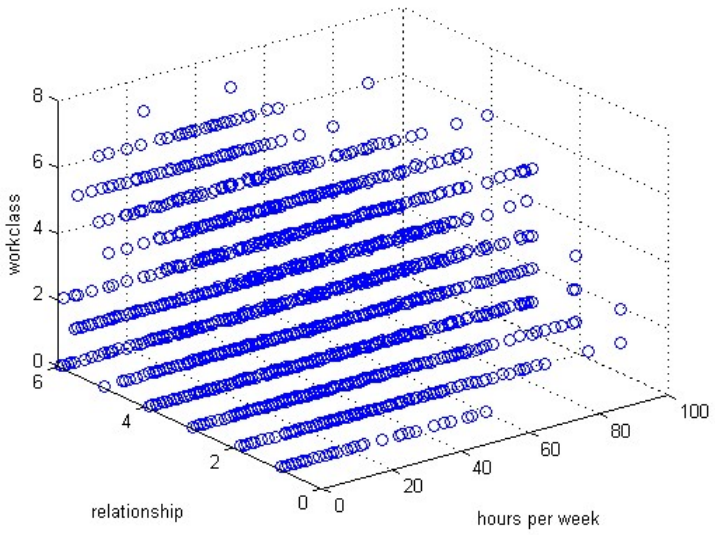

Figure 4.8: Scatter plot of relationship vs workclass vs hours-per-week(ADULT)

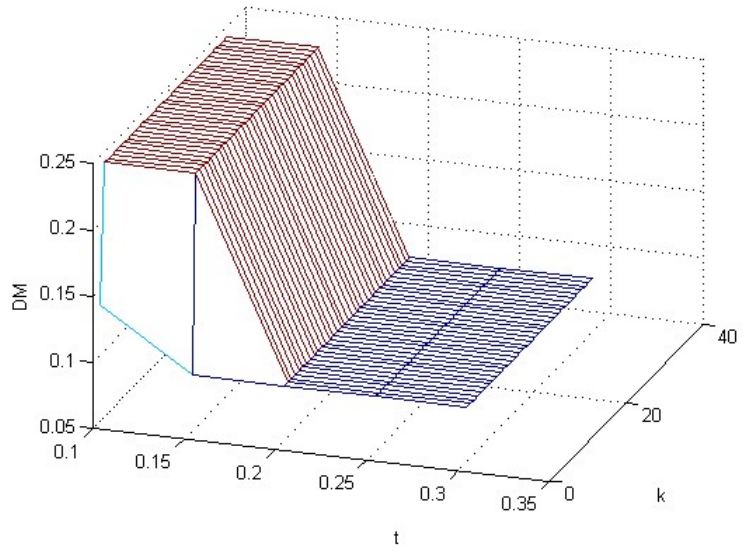

Figure 4.9: t vs k vs Discernibility Metric

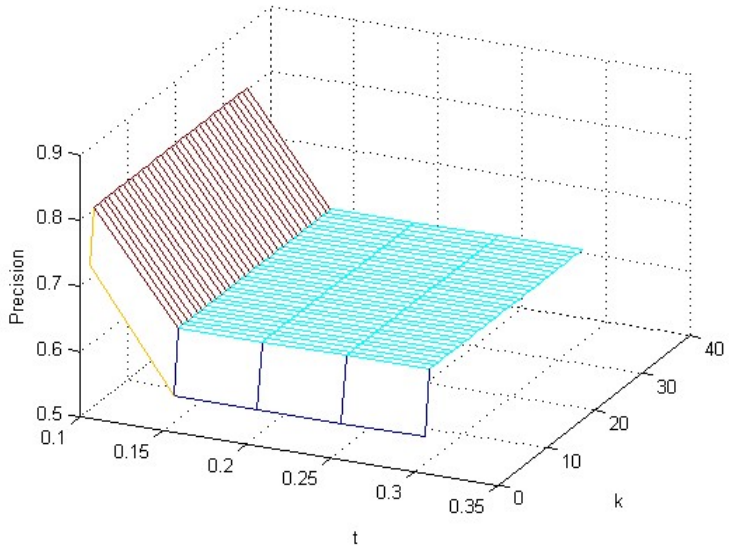

Figure 4.10: t vs k vs Precision

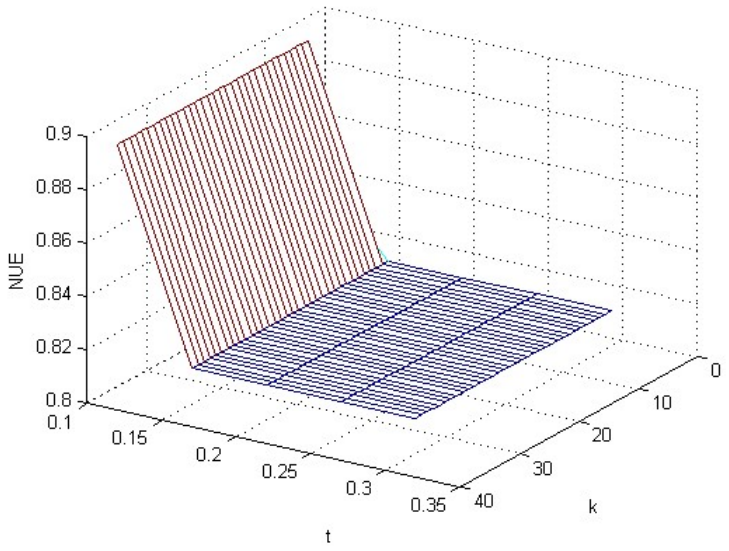

Figure 4.11: t vs k vs Non-Uniform Entropy

Chapter 5

Conclusion

The proposed scheme utilizes the Multiple Sensitive Attributes in any given dataset to determine t. The experiments have been conducted on the well-known dataset ADULT [22].The first experiment was finding Niche Overlap values which gave the result to be $t = 0.21$. Already in [6] the value *0.20* is declared to be the point at which least data degredation takes place. To further bolster the claim made in this thesis the following experiment was done.

The verification for the bound was done using the Flash Anonymization toolbox [23],to obtain the values of t which correspond to good bounds on the actual t-values which are used for anonymization purposes.In Figures 4.9,4.10,4.11 where all the Information Loss metrics namely *Discernibility Metric, Non-Uniform Entropy* and *Precision* are plotted against t and k , show that upto the bound of t demarcated by us the Information Loss gradually decreases and then becomes constant or increases. This shows that if we want to achieve anonymization with minimum Information Loss the bound acts as an upper limit and there is no need to run the anonymization algorithm after that value of t.

The effectiveness of our method can be measured in terms of time saved while determining t for yet unknown datasets which have to be anonymized. Being an NP-Hard problem with an average dataset size of $\approx$ 30000($ADULT$) every anonymization instance takes a few minutes [6] which can be entirely saved.

The datasets used here consist of similar number of partitioning classes for Sensitive Attributes which make the job easier in terms of finding Niche Overlap [7]. If the classes of SA are not similar in number then we can use the Domain Generalization Hierarchy can be used to first generalize to a certain level before applying the proposed algorithm.

All said and done it has to be accepted that if any dataset has only one sensitive attribute as decided by the data disseminating authority then the proposed method for determining t cannot be employed to determine t for that particular case.

5.1 Future Scope

The work carried on here can be extended to all known datasets like FARS, IRIS etc. so as to provide a generic exploration of the scheme proposed here.This can help with presenting relevant details like values of t and k which give anonymization and minimum data distortion on a public platform for future use by researchers.

Also, there is a need to provide tolerance values regarding the approximation of t, when dealing with sensitive attributes which differ in number of partitioning classes. This can greatly reduce the work in terms of time spent for anonymization for matching number of classes would become zero.

Dissemination

Debaditya Roy, Sanjay Kumar Jena.

Determining t in t-closeness using Multiple Sensitive Attributes.

International Journal of Computer Applications(accepted),May 2013

Bibliography

[1] L. Sweeney. K-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge-based systems*, 10(5):557–570, 2002.

[2] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART (PODS)*, page 188, New York, 1998. ACM.

[3] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.

[4] D. Kifer A. Machanavajjhala, J. Gehrke and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[5] D. Kifer A. Machanavajjhala, J. Gehrke and M. Venkitasubramaniam. Privacy: Theory meets practice on the map. In *Proceedings of the 24th IEEE International Conference on Data Engineering(ICDE)*, pages 277–286. IEEE, 2008.

[6] T. Li N. Li and S. Venakatasubramaniam. t-closeness: Privacy beyond k-anonymity and t-closeness. In *Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE)*. IEEE, 2007.

[7] K.C. Burns S.W. Geange, S. Pledger and J.S. Shima. A unified analysis of niche overlap incorporating data of different types. *Methods in Ecology and Evolution*, 2(2):175–184, 2011.

[8] H. Liang and H. Yuan. On the Complexity of t-closeness Anonymization and Related Problems. arXiv preprint arXiv:1301.1751 2013.

[9] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[10] S. Venkatasubramaniam. Closeness: A New Privacy Measure for Data Publishing. *IEEE Transaction on Knowledge and Data Engineering*, 22:943–956, 2010.

[11] C. Tomasi Y. Rubner and L.J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[12] M.S. Nikulin. *Encyclopedia of Mathematics*, volume ISBN 978-1556080104, chapter Hellinger Distance. Springer, 2001.

[13] S.K. Jena P. Khaitan, S.B. Korra and B. Majhi. Approximation algortihms for optimizing privacy and utility. In *Proceedings of 2nd International Conference on Computer Science and its Applications CSA,South Korea*, pages 59–64, 2009.

[14] A. Bhattacharya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of Calcutta Mathematical Society*, 35(MR0010358):99–109, 1943.

[15] Z. Chen T.S. Gal and A. Gangopadhyay. A Privacy Protection Model for Patient Data with Multiple Sensitive Attributes. *International Journal of Information Security and Privacy*, 2(3):28–44, 2008.

[16] C. Wang D. Lv Y. Ye, Y. Liu and J. Feng. Decomposition: Privacy preservation for Multiple Sensitive Attributes. In *Database Systems for Advanced Applications*, pages 486–490. Springer, 2009.

[17] D. Das and D.K. Bhattacharya. Decomposition+: Improving l-diversity for Multiple Sensitive Attributes. In *Advances in Computer Science and Information Technology, Computer Science and Engineering*, pages 403–412, Berlin Heidelberg, 2012. Springer.

[18] R.N.V. Vishnu Murthy M.V.R. NarasimhaRao, J.S. VenuGopalkrisna and C.R. Ramesh. Closeness: privacy measure for data publishing using multiple sensitive attributes. *International Journal of Engineering Science and Advanced Technology*, 2(2):278–284, March-April 2012.

[19] R.J. Bayardo and R. Agarwal. Data Privacy through Optimal k-anonymization. In *Proceedings of the 21st IEEE International Conference on Data Engineering(ICDE)*, pages 217–228. IEEE, 2005.

[20] C.E. Shannon. *The Mathematical Theory of Communication.* University of Illinois Press, 1949.

[21] R. Chen B.C. M. Fung, K. Wang and P.S. Yu. Privacy-preserving data publishing: A Survey of Recent Developments. *ACM Computing Surveys*, 42:1–53, 2010.

[22] UCI. Uci machine learning repository. http://www.ics.uci.edu/mlearn/mlrepository.html.

[23] F. Kohlmayer and F. Praer. Arx - Powerful Data Anonymization. http://arx.deidentifier.org/, 2013.